applied sciences

MDPI

Article

A Methodology for Advanced Manufacturing Defect Detection through Self-Supervised Learning on X-ray Images

Eneko Intxausti [1,*], Danijel Skočaj [2], Carlos Cernuda [1] and Ekhi Zugasti [1]

[1] Electronics and Computer Science, Mondragon Unibertsitatea, Loramendi 4, 20500 Mondragon, Spain
[2] Faculty of Computer and Information Science, University of Ljubljana, Večna Pot 113, 1000 Ljubljana, Slovenia
* Correspondence: eintxausti@mondragon.edu

Abstract: In industrial quality control, especially in the field of manufacturing defect detection, deep learning plays an increasingly critical role. However, the efficacy of these advanced models is often hindered by their need for large-scale, annotated datasets. Moreover, these datasets are mainly based on RGB images, which are very different from X-ray images. Addressing this limitation, our research proposes a methodology that incorporates domain-specific self-supervised pretraining techniques using X-ray imaging to improve defect detection capabilities in manufacturing products. We employ two pretraining approaches, SimSiam and SimMIM, to refine feature extraction from manufacturing images. The pretraining stage is carried out using an industrial dataset of 27,901 unlabeled X-ray images from a manufacturing production line. We analyze the performance of the pretraining against transfer-learning-based methods in a complex defect detection scenario using a Faster R-CNN model. We conduct evaluations on both a proprietary industrial dataset and the publicly available GDXray dataset. The findings reveal that models pretrained with domain-specific X-ray images consistently outperform those initialized with ImageNet weights. Notably, Swin Transformer models show superior results in scenarios rich in labeled data, whereas CNN backbones are more effective in limited-data environments. Moreover, we underscore the enhanced ability of the models pretrained with X-ray images in detecting critical defects, crucial for ensuring safety in industrial settings. Our study offers substantial evidence of the benefits of self-supervised learning in manufacturing defect detection, providing a solid foundation for further research and practical applications in industrial quality control.

Keywords: defect detection; manufacturing; optical quality control; deep learning; self-supervised learning

check for updates

Citation: Intxausti, E.; Skočaj, D.; Cernuda, C.; Zugasti, E. A Methodology for Advanced Manufacturing Defect Detection through Self-Supervised Learning on X-ray Images. *Appl. Sci.* **2024**, *14*, 2785. https://doi.org/10.3390/app14072785

Academic Editor: Andrzej Sioma

Received: 22 February 2024
Revised: 17 March 2024
Accepted: 20 March 2024
Published: 26 March 2024

1. Introduction

Over the years, extensive research has been conducted to enhance the visual inspection of manufacturing products using X-ray imaging. The main focus has been on developing automated processes capable of identifying defective products. Actually, manual analysis of each piece is not only a repetitive and fatiguing task for operators, but their accuracy can also tend to decrease over time [1]. In contrast, data-driven approaches not only ensure uniform performance over prolonged periods but also effectively mitigate the risk of human errors. Therefore, they can significantly aid in the operators' decision-making process.

Recent advancements in deep-learning-based approaches have emerged as the leading solution for a variety of tasks across multiple domains [2]. Specifically in the context of manufacturing defect detection, these techniques are now considered state of the art, significantly outperforming traditional methods [3–5]. However, their performance is conditioned by a significant condition: their data-hungry nature. These approaches require extensive datasets of labeled images during training to learn effective visual representations. Consequently, their effectiveness can drastically diminish when only a limited number of images is available, highlighting a critical challenge in their application. This challenge becomes even more notable in defect detection as the acquisition of extensive, accurately labeled datasets proves especially difficult within industrial environments.

A potential solution to mitigate the data requirements is the adoption of transfer learning. This approach fundamentally involves leveraging knowledge acquired from broadly defined problems and applying it to more specific, targeted tasks. For vision models, pretraining on extensive datasets like ImageNet [6] enhances their capability for various downstream tasks. However, despite the demonstrated effectiveness of these models in manufacturing defect detection, the performance can be impacted by (1) the disparity between ImageNet images and X-ray images and (2) the potential bias of these models towards the specific categories present in the dataset.

However, the need for a labeled dataset in these pretraining approaches often presents a bottleneck in the development of deep learning models. Self-supervised learning emerges as a solution, enabling the extraction of significant features from images without relying on labeled data and thus making efficient use of the vast volumes of available data. In the context of manufacturing defect detection, self-supervised approaches allow for the utilization of large datasets of unlabeled X-ray images, common in industrial settings. This facilitates the training of a model adept at understanding X-ray image representations and reliable in extracting the most relevant features. Subsequently, the model becomes a suitable backbone for downstream tasks that involve a labeled X-ray dataset, enhancing its applicability and effectiveness in precise defect identification.

The main contribution of this work is the introduction of a novel methodology for defect detection in manufacturing parts where we utilize specialized pretraining approaches on X-ray images to develop models with enhanced feature extraction capabilities. By implementing a pretraining phase with images from the specific manufacturing field, our models evolve into highly adept feature extractors, showing a significant improvement in identifying task-relevant features compared to models initialized with ImageNet weights. This methodology is validated by the strong performance of these models in defect detection tasks within manufacturing contexts, as seen through evaluations both on a public benchmark dataset and in real-world industrial environments. Such validation underscores the practical efficacy and significance of our approach in scenarios where precise defect identification is crucial. Through this work, we contribute to the manufacturing defect detection domain by showcasing how specialized pretraining can effectively extract relevant features from X-ray images, thus offering new insights into applying this technology in industrial environments.

2. Related Work

2.1. Manufacturing Defect Detection

Numerous efforts have been made over the years to automate the task of defect detection in manufacturing using computer vision methods. Currently, manufacturing defect detection is a well-established problem in the field of computer vision and has been widely applied in numerous industrial quality control processes [7–9]. Initial attempts in defect detection used image comparison [10] and Fourier Transform [11] to identify defects, followed by statistical techniques for feature extraction [8,12]. These features were then classified using machine learning to distinguish between defective and non-defective items. However, this manual feature extraction was product specific and not universally applicable, limiting the transferability of learned knowledge.

The emergence of deep-learning-based approaches has improved the accuracy in manufacturing quality control [3–5,13–15]. Their complex structures are able to retain and automatically learn the information contained in the image, more effectively facilitating the image processing compared to previous techniques. Deep-learning-based models are built in an end-to-end manner so handcrafting processes are not required to extract discriminant features [3,16]. In fact, the feature extraction process is carried out automatically from raw images, followed by a classifier head that learns the boundary between defective and non-defective features.

In recent years, numerous studies have employed deep Convolutional Neural Networks (CNNs) for image-level classification, effectively differentiating between defective and non-defective manufacturing images [1,8]. These works focused on refining the feature extraction process, considering it crucial for developing an effective defect detector. Kuo et al. [1] compared different feature extraction backbones on surface images with sandblasting defects. Additionally, Wang et al. [17] used a self-attention module to extract and classify features from small defects.

However, the aim of defect detection is not only to classify a product at image level but also to locate the defects throughout the image when the product is defective. Therefore, research shifted towards the application of object detection models to incorporate location information [3,13–15,18]. Ferguson et al. [3] applied Faster R-CNN [19] to a GDXray benchmark dataset [20], obtaining satisfactory results. Soon after, they improved their results by applying a Masked R-CNN [4] pretrained on the COCO dataset [21]. Du et al. [13] proposed several improvements to the Faster R-CNN for defect detection in automobile manufacturing products, including the Feature Pyramid Network [22] (FPN), RoIAlign [23], and data augmentation techniques. Furthermore, Wang et al. [17] were the first to integrate self-attention mechanisms with CNNs, enhancing image defect detection by extracting subtle features from general features.

The main challenge facing deep-learning-based approaches is their reliance on a large volume of images to achieve reliable performance [24]. Indeed, applying transfer learning with pretrained models offers a suitable initial step [4]. However, there is a notable disparity between X-ray images and those typically used in ImageNet pretraining; X-ray images are grayscale and depict specific manufacturing parts, contrasting with the RGB images of varied everyday objects and scenes from ImageNet. This discrepancy leads to a significant domain shift.

Although there have been significant advancements in applying defect localization to manufacturing defect detection, a notable gap exists in the use of self-supervised learning on X-ray manufacturing images. To the best of our knowledge, this specific approach remains unexplored, representing a promising direction for the enhancement of the accuracy and efficiency of defect detection in industrial settings. We hypothesize that self-supervised learning, specifically applied to manufacturing X-ray images, could be an effective strategy for overcoming the reliance on large, labeled datasets in deep learning. By adopting this method for pretraining, our goal is to extract and refine the most representative features intrinsic to this specific domain, which could subsequently enhance the performance of downstream defect detection tasks. Therefore, this advancement would increase the effectiveness and reliability of defect detection in real-world industrial applications.

2.2. Self-Supervised Learning

As has been mentioned, it is widely acknowledged that current deep learning algorithms require large-scale training datasets to learn intrinsic data representations and reach satisfactory generalization ability. In supervised learning, labeled data are compulsory during the training, so large datasets have to be annotated. This can be seen as a bottleneck because of the time-consuming and expensive labeling process, as well as infeasible in several fields. Moreover, the models trained in a supervised way are heavily dependent on manually annotated labels.

Currently, self-supervised learning is a popular alternative to learning visual representations of images without annotated data. It leverages the huge number of unlabeled data available to train a model solving different pretext tasks. During this process, the model learns inherent image features that can be used for several final purposes [25]. As proposed in the survey of Liu et al. [26], self-supervised learning approaches can be summarized as generative based [27], contrastive based [28,29], or GAN based [30]. The survey shows that generative- and GAN-based approaches underperform in classification tasks compared to contrastive-based approaches. Therefore, as defect detection can be seen as a particular classification problem, our analysis is focused on the self-supervised contrastive-based techniques.

The main idea of the contrastive approach is to build a simple discrimination problem based on a pretext task that helps the model to learn representative features from images. The pretext task clusters the images into different groups under the assumption that images from the same group are semantically similar whereas images from different groups are not [25]. According to this, contrastive learning tries to minimize the distance between image features from the same groups (known as *positive pairs*), thus decreasing intra-class similarity, and maximize the distance between image features from different groups (*negative pairs*), increasing inter-class similarity.

Attempting to define the pretext task, Wu et al. [31] introduced instance discrimination with a memory bank for efficient training, emphasizing the importance of a large number of negative samples for improved performance. This concept was further refined by He et al. [28], who used a momentum encoder and data augmentation to generate positive pairs from the same image. Chen et al. [29] proposed SimCLR, eliminating the need for a memory bank and using current batch negative samples for contrastive loss calculation, although requiring large batch sizes for better representation quality. Subsequent studies focused on the impact of hard negatives and batch size balance [32,33]. Grill et al. [34] addressed the large batch size issue by relying solely on positive pairs and enhancing the momentum encoder, offering robustness in smaller batch scenarios. Yet, the SimSiam model by Chen et al. [35] represents a leap forward with its simpler architecture and unique stop-gradient feature, effectively performing without needing negative samples, large batches, or momentum encoders.

In the realm of self-supervised learning, a significant evolution has been observed with the integration of Transformers into computer vision, moving away from the traditional reliance on CNN-based approaches [36]. These models, well known for their ability to capture complex relational dependencies within image sequences, have shown remarkable scalability in handling large-scale networks and datasets [37]. This advancement led to the development of hierarchical self-attention architectures focused on extracting multi-resolution features essential for detailed applications like defect detection [38–40].

This transition towards Transformers in vision aligns seamlessly with the broader advancements in self-supervised learning. Vision Transformers, paralleling NLP techniques like BERT's token masking [41], apply similar pretraining strategies. By masking parts of images and reconstructing them, this method enables learning from large, unlabeled datasets, reducing dependence on annotated data [42]. This reflects unified progression in exploiting unlabeled data for complex visual tasks, emphasizing the importance of robust visual representations [43].

3. Materials and Methods

This section provides a detailed overview of the methodology we employed to achieve our main goal: improving defect detection performance in a real-world industrial environment using a pretraining stage. We detail and compare the two self-supervised techniques used in our approach, emphasizing how each contributes to enhancing the learning of visual representations. Additionally, we discuss their potential impact on improving defect identification within industrial settings.

3.1. Our Methodology

Our methodology is centered on the application of self-supervised learning techniques to construct defect detectors for manufacturing with the goal of enhancing the current state of the art in manufacturing defect detection. It stands out from previous approaches in manufacturing defect detection by exploiting large datasets of unlabeled X-ray images, which are commonly found in industrial settings, to train defect detectors in a self-supervised manner.

The pretraining phase of our methodology involves the development of models capable of automatically learning and extracting critical visual features from X-ray images, independent of annotated data. In this stage, the models are trained in a self-supervised way on an unlabeled dataset of X-ray manufacturing images. In the subsequent sections of this paper, we comprehensively detail two distinct self-supervised learning techniques. These methods have been specifically chosen for their demonstrated efficiency and superior performance in prior pretraining tasks. As a result of the training, we leverage the inherent characteristics and patterns present in these images, thus enabling the models to identify subtle yet significant anomalies that signal defects. Additionally, this pretraining stage is not burdened with high data gathering costs. The significant expense in data preparation usually arises from the need for precise image annotations rather than the collection of raw data itself. In our approach, raw images can be efficiently collected through an X-ray camera integrated into the production line, enabling the acquisition of a comprehensive set of manufactured products.

After the self-supervised pretraining, the model serves as a refined feature extractor for specific downstream tasks, particularly for defect detection in manufacturing. This approach takes full advantage of the improved ability of the model to understand and interpret the characteristics of X-ray images. Subsequently, in line with established methods in the literature [3,4,13], we integrate the feature extractor into an object detection framework. The model undergoes fine-tuning on a carefully annotated target dataset, where defects are delineated using bounding boxes or pixel-wise annotations, clearly distinguishing them from the image background. The emphasis during this fine-tuning stage is on training the detector to accurately localize and identify defect features within the entire image, enhancing the precision of the defect detection. Diverging from traditional methods that commence fine-tuning with ImageNet weights, our methodology employs features pertinent to manufacturing images, derived from the self-supervised pretraining. This approach potentially enables the model to assimilate local defect features more effectively as it integrates the learned visual representations with traditional defect detection mechanisms, leveraging the global representations of manufacturing products acquired during the pretraining phase. As a result, the final defect detection system is expected to be more robust and reliable in detecting defects, offering a significant improvement over traditional approaches.

Figure 1 shows an overview of the methodology, which encompasses two key stages: pretraining and fine-tuning. In the pretraining phase, the backbone of the model is trained through a self-supervised learning approach. This training equips the model with the ability to specifically extract features from X-ray manufacturing images. Following this, the trained backbone is adapted into an object detection framework. Here, it undergoes a fine-tuning process with an annotated target dataset, which further refines its defect detection capabilities for manufacturing images. To assess the effectiveness of our methodology, we have chosen two self-supervised learning approaches: SimSiam [29] and SimMIM [44]. Detailed descriptions of these techniques are provided in the following subsections.

3.2. SimSiam

We employed SimSiam [29] for pretraining a feature extractor specifically tuned to discriminate between relevant features from manufacturing images. SimSiam was chosen for its optimal trade-off between simplicity and effectiveness in image classification tasks. Its advantage lies in not requiring large batch sizes or a momentum encoder for training, making it feasible to train on multiple GPUs without demanding significant resources.

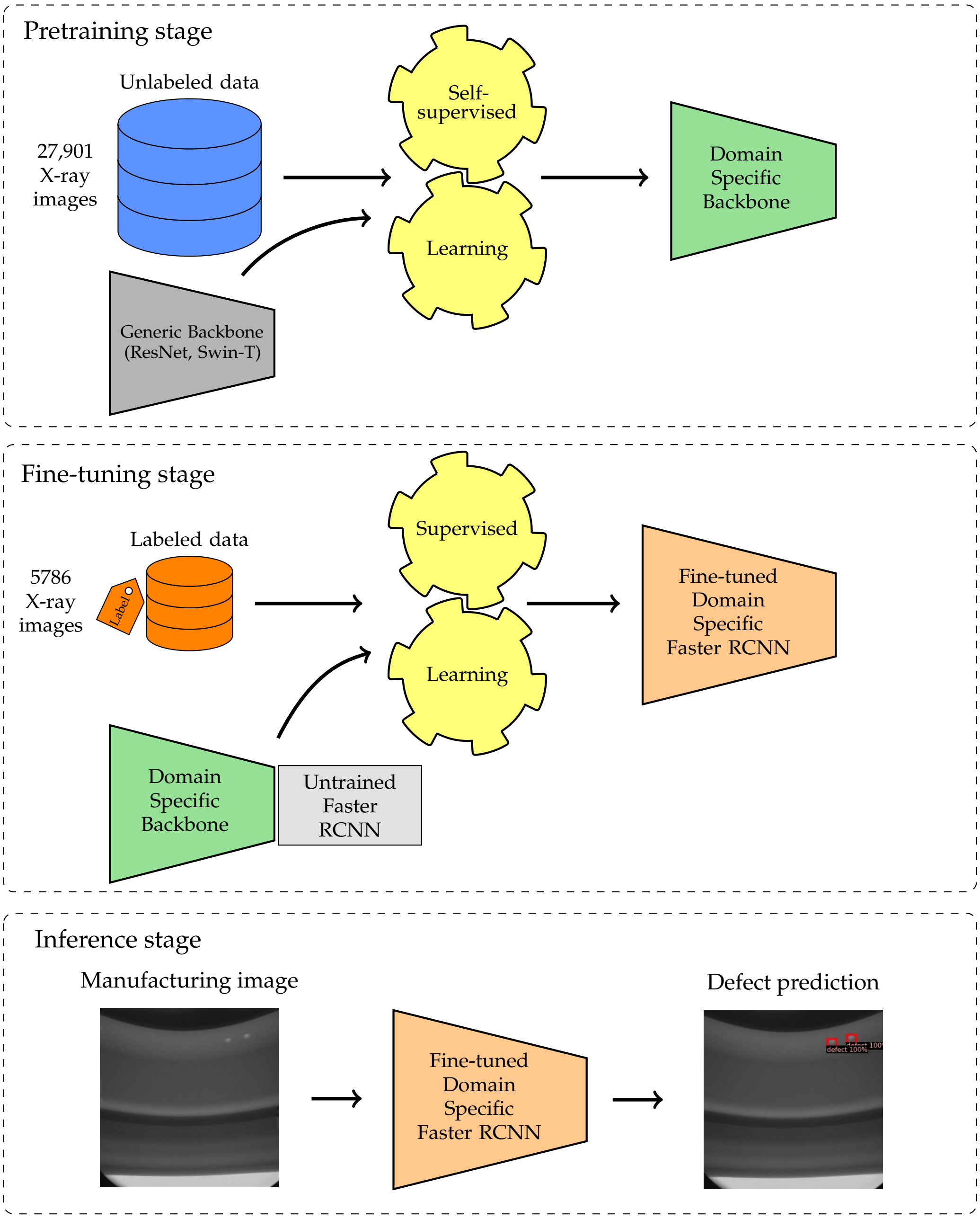

Figure 1. Methodology overview: The process comprises two stages, pretraining and fine-tuning. Initially, we utilize self-supervised learning techniques SimSiam and SimMIM to train the generic backbone on large volumes of unlabeled X-ray images. This initial phase helps the model to capture critical visual features relevant to manufacturing defects, eliminating the need for data annotations. Subsequently, the pretrained backbone is integrated into an object detection framework, where it undergoes fine-tuning with an annotated target dataset. The final inference stage applies the fine-tuned model to accurately detect defects in manufacturing images, using the robust features learned in earlier phases for improved precision and reliability.

SimSiam is a contrastive learning approach that employs a Siamese-network-based architecture. This method trains the model to maximize the agreement between different augmented views of the same data using contrastive loss in the latent space. It generates two random augmentations, x_1 and x_2, from an image x which are then processed by an encoder network f. This encoder f, comprising a backbone such as ResNet50 [45] and a projection multilayer perceptron (MLP) head, shares its weights across views for consistent feature extraction. Additionally, a separate MLP, denoted as h, serves as a prediction head, transforming one view's output to align with the other view. Following the notation of the original paper, the transformed and original vectors are represented as $p_1 \triangleq h(f(x_1))$ and $z_2 \triangleq f(x_2)$, respectively. In a symmetrized manner, the model also processes the augmentations inversely, producing $z_1 \triangleq f(x_1)$ and $p_2 \triangleq h(f(x_2))$. This symmetrization in the loss function significantly enhances the robustness and learning effectiveness of the model. The core objective of SimSiam is to minimize the negative cosine similarity between pairs (p_1 and z_2 and p_2 and z_1), thereby aligning these representations despite their augmentation-induced variations.

$$\begin{aligned} \mathcal{D}(p_1, z_2) &= -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2} \\ \mathcal{D}(p_2, z_1) &= -\frac{p_2}{\|p_2\|_2} \cdot \frac{z_1}{\|z_1\|_2} \end{aligned} \tag{1}$$

where $\|p_1\|_2$ and $\|p_2\|_2$ are the L2 norms of p_1 and p_2, respectively. A pivotal element of the SimSiam approach is the stop-gradient operation, which is key to preventing a collapsing solution where all outputs converge to the same vector. This operation is applied to the output of the encoder f, treating it as a constant during a portion of the training. This technique helps maintain diversity in the learned features, ensuring that the model does not trivially minimize the loss by converging to a constant output, as illustrated in Figure 2. The final form of the loss function is defined as follows:

$$\mathcal{L} = \frac{1}{2}\mathcal{D}(p_1, stopgrad(z_2)) + \frac{1}{2}\mathcal{D}(p_2, stopgrad(z_1)) \tag{2}$$

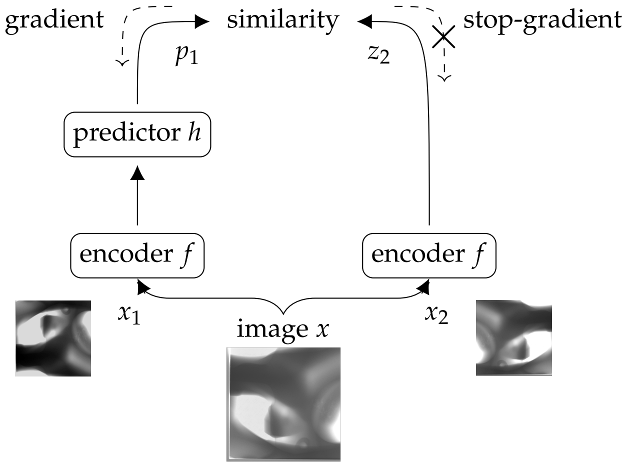

Figure 2. Overview of the SimSiam architecture [29]. From an original image, two augmented versions are processed by a Siamese network to extract and align features, utilizing a stop-gradient operation to prevent collapsing solutions. This strategy minimizes negative cosine similarity, enhancing consistent feature identification across variations with efficient contrastive learning.

The goal is for the model to learn representations that are invariant to the augmentations, ensuring the model yields consistent outputs for different views of the same image. This necessitates the careful selection of augmentations that are sufficiently invariant for trivial variations to be ignored, but not excessively, so important information for downstream tasks is not discarded [46]. Achieving this balance enables the model to effectively identify and highlight relevant features. In the original paper, the authors employed a range of augmentations, including geometric (such as cropping the image by up to 20% and horizontal flipping), color modifications (brightness, contrast, saturation, hue, and grayscale), and Gaussian blurring. We followed the same strategy in our approach.

3.3. SimMIM

Transformers have emerged as state of the art for a wide range of vision-related tasks, primarily due to their exceptional capability in feature extraction and representation learning [37]. This progress inspired our exploration into their application for defect detection in manufacturing processes.

Furthermore, with the aim of maximizing the capabilities of the Transformers, the selection of a pretraining approach is critical [47]. Considering the need for defect detection models to accurately detect features across multiple resolutions, we selected the SimMIM approach [44] for the pretraining phase on X-ray manufacturing images. This choice was informed by the compatibility of SimMIM with multi-scale Vision Transformers, which is essential for effective feature discernment at various scales. Adopting the masked image modeling approach, SimMIM aims to acquire visual representations by reconstructing randomly masked image patches. This approach allows for a deeper understanding of the underlying visual context.

The SimMIM approach incorporates an encoder–decoder strategy. The encoder is responsible for extracting latent feature representations from the unmasked sections of image patches. The decoder reconstructs the pixel values of masked patches. l_1 loss is used in order to train this model. This process involves down-sampled-resolution feature maps from the encoder being mapped back to original-resolution feature maps through a 1 × 1 convolution layer. The l_1 loss function is then applied specifically to the masked pixels using the following formula:

$$\mathcal{L} = \frac{1}{\Omega(x_M)} \|\mathbf{y}_M - \mathbf{x}_M\|_1 \quad (3)$$

where x and y represent the input image and predicted values, respectively, with M indicating masked pixels and Ω the count of images. Figure 3 provides a schematic representation of the model architecture.

The key aspect of SimMIM is its efficient reconstruction process, which demonstrates that the lightweight prediction head can perform comparably, if not better, than more complex reconstruction decoders [44]. This efficiency not only maintains accuracy but also significantly reduces pretraining time and computational resource demands.

The masking strategy in SimMIM, crucial for maximizing its effectiveness, involves applying random masking to the image. SimMIM demonstrates that either larger patch sizes or higher masking ratios lead to improved performance in downstream tasks [44]. This is attributed to the reduced correlation between unmasked and masked pixels when larger image areas are masked, compelling the model to learn more semantic information instead of replicating the closest highly correlated pixels. For instance, in ImageNet classification, SimMIM achieves optimal results with a 32-pixel masking patch size and a 60% masking ratio as it forces the model to infer more complex and less obvious image features.

A notable advantage of the SimMIM strategy over approaches like SimSiam is that it does not require specifically chosen data augmentation for each problem. This characteristic simplifies the application of the method across various scenarios, reducing the need for customized augmentation strategies and thereby optimizing the process of model training and adaptation to different tasks.

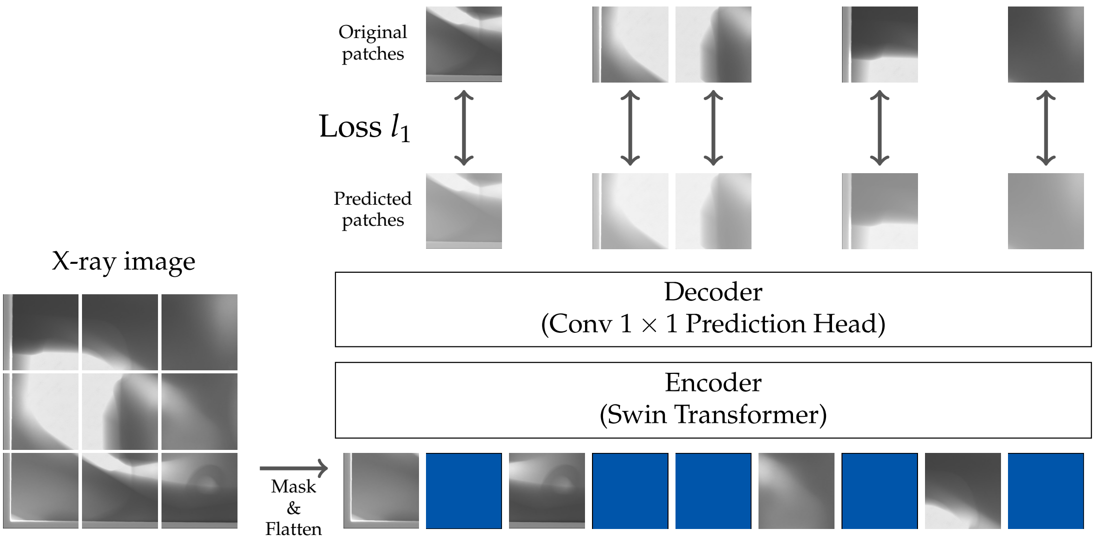

Figure 3. Overview of the SimMIM architecture [44]. SimMIM is based on an encoder–decoder mechanism used for feature extraction and the reconstruction of masked image patches. The encoder processes unmasked parts of the image, while the decoder aims to reconstruct the masked areas. It is trained using an l_1 loss function to map down-sampled-resolution feature maps back to their original resolution.

4. Implementation and Results

In this section, we present the experiments conducted to evaluate the effectiveness of the models pretrained on X-ray images of manufacturing products. We detail the specific datasets employed in our experiments and discuss the methodologies applied. This section also discusses the results obtained, offering insights into the performance implications of different pretraining approaches.

4.1. Pretraining on X-ray Images

We utilized the two above-mentioned self-supervised methods, SimSiam [35] and SimMIM [44], to carry out a model pretraining stage on X-ray manufacturing images. Our hypothesis was that self-supervised pretraining on a large dataset of X-ray images could improve the performance of models in manufacturing defect detection downstream tasks.

For the pretraining phase, a considerable volume of images is essential. However, the self-supervised approach eliminates the need for annotations, making it easier to gather a large unlabeled dataset. In industrial settings where images without annotations are readily available, this approach is particularly advantageous, circumventing the costly and labor-intensive process of manually annotating defects. Later, an in-depth analysis of the datasets utilized in our study is provided in Section 4.2.

To assess the impact of pretraining on X-ray images, we contrasted models fine-tuned from ImageNet with those fine-tuned after X-ray image pretraining. We employed two evaluation protocols for comparison: linear classification and fine-tuning.

- **Linear classification**: This evaluation method is a widely recognized protocol for assessing the efficacy of self-supervised learning techniques [28,29,35,48]. In this approach, the pretrained backbone network is kept fixed (frozen), and the assessment focuses on training a singular fully connected layer which is appended to the top of the backbone. This additional layer, incorporating a softmax activation function, processes the feature vector derived from the global average pooling layer of the backbone, and it is trained on a target dataset. The primary objective of this protocol is to evaluate and compare the classification capabilities of different backbone networks in a standardized setting, thus providing insights into their relative performance and adaptability;

- **Fine-tuning**: In this protocol, the features of the pretrained backbone are adapted for specific downstream tasks, such as defect detection. This process involves using the pretrained weights of the backbone as an initial setup, which is then fine-tuned for task-specific architectures like Faster R-CNN [19] and UNet [49] on a target dataset, enhancing their performance in specialized applications.

Fine-tuning generally surpasses linear classification in performance, which is largely attributable to its utilization of more intricate architectures and the inherent constraints associated with employing a frozen backbone in the linear classification approach. By concurrently analyzing both protocols, we facilitated a quick yet comprehensive evaluation of various feature extractors. This dual-method analysis offers a reliable and robust framework for assessing models, aligning particularly well with the demands of real-world industrial applications where practical performance and adaptability are paramount. Nonetheless, the primary objective of this study was to enhance the effectiveness of existing manufacturing defect detection systems. In this context, fine-tuning emerges as the more critical metric given its direct impact on improving practical detection capabilities. Consequently, we chose fine-tuning as our principal protocol for manufacturing defect detection.

4.2. Use Cases

In the following experiments, we aimed to assess the efficacy of the methods using two main datasets.

4.2.1. Industrial Dataset

The industrial dataset comprised X-ray manufacturing images obtained from a production line within a manufacturing company. The experimental data were obtained using an X-ray machine that captures images of manufacturing products on the production line. These images were stored and later annotated by an X-ray expert for use in model training.

The acquisition of X-ray images employed in this study was subject to the proprietary and confidential protocols required by the collaborating manufacturing company. It is important to note that our research only began after the images had been acquired, focusing solely on the deep-learning-based analysis of these images. We did not participate in or have access to the image acquisition process. Highlighting this fact is essential for a clear understanding of our research focus, which was solely analyzing the provided images.

This dataset was divided into two subsets: an extensive pretraining dataset and a smaller, annotated target dataset for model evaluation and fine-tuning. Initially, we compiled a substantial set containing 27,901 unlabeled images, which we designated as the pretraining dataset. This dataset served as the foundation for the pretraining stage, primarily due to its lack of annotations. The acquisition of unlabeled images in a realistic industrial setting is relatively straightforward. Subsequently, a second set was developed with image-level and ground-truth annotations, consisting of 5786 images. This set was smaller than the first, primarily because annotating images is resource intensive. However, this size was deemed appropriate for our purposes. We designated this set as the target dataset, which was utilized both for evaluating the backbone using the linear classification protocol and for fine-tuning the model specifically for manufacturing defect detection in an industrial environment. In both datasets, grayscale images depict various perspectives of the same product, each with dimensions of 1024 × 1024 pixels. In the target dataset, ground-truth annotations for all defects are provided as bounding boxes, ensuring precise reference points for defect detection. The target dataset contains 19 annotated defect types of diverse sizes, scales, and intensities, exhibiting significant variability. Following a preprocessing step, defects are categorized into two main groups: critical and minor. As their names imply, the accurate detection of critical defects is strictly necessary due to potential safety implications. Detecting minor defects is also crucial, although their severity is comparatively lower, rendering their identification equally important but less critical for safety concerns. The defect distribution includes 7 critical defect types and 12 minor defect types, posing a challenging defect detection problem with a dual objective: locating defects on the product and classifying them into one of the two categories. Concerning the dataset distribution, a significant class imbalance was evident, with 1784 samples for critical defects and 4002 samples for minor defects. The partitioning of the target dataset into training and test sets while maintaining class proportions is detailed in Table 1.

Table 1. Distribution of images for critical and minor defects in the target dataset.

	Critical Defects	**Minor Defects**
Train	1424	3200
Test	360	802

4.2.2. GDXray Dataset

As the above-mentioned dataset was privative, we also considered evaluating the methods on a benchmark dataset. GDXray [20] is a publicly available dataset of X-ray images. It contains 19,407 images organized into five groups: castings, welds, baggage, natural objects, and settings, each with multiple series. As we were dealing with manufacturing defect detection, we focused on the castings group, which is related to manufacturing and which comprised 2727 X-ray images from 67 series primarily featuring automotive parts like aluminum wheels and knuckles. These images are annotated with bounding boxes, similar to in the first dataset, offering a benchmark for assessing our methods in manufacturing defect detection. However, it is noteworthy that, unlike our industrial target dataset, GDXray does not categorize defects into different types. Additionally, considering that the existing literature utilizing GDXray primarily addresses defect detection rather than classification, we opted to refrain from performing the linear classification evaluation for this dataset. Consequently, our analysis of the GDXray dataset was confined to fine-tuning evaluation.

4.3. Results

In this section, we present several experiments to evaluate the advantages of applying a specific pretraining stage for X-ray manufacturing images.

4.4. Linear Classification

Initially, we trained a backbone model using both SimSiam and SimMIM. We followed the default scheme from original papers in both cases, and we trained the models for 100 epochs. Afterwards, we implemented the standard evaluation protocol for the pretrained models; we added a linear layer at the top of the model, with the rest of the layers of the backbone frozen. We trained this classification model on the target dataset with critical and minor defects. We established a baseline using a model with ImageNet pretrained weights. For SimSiam, we used ResNet50 [45] as the feature extractor, generating a 2048-dimensional feature vector for the classification layer. Conversely, for SimMIM, we used the original Swin Transformer [38] configured with a 12-pixel window size. The feature vector in the case of Swin Transformer was 1024 dimensional. The results of this linear classification on our target dataset are detailed in Table 2.

Table 2 compares the performance of different backbone architectures using ImageNet weights versus weights obtained through self-supervised pretraining methods with X-ray images. From the results, we observed that, for the ResNet backbone, the SimSiam pretraining method outperforms the supervised approach in both Average Precision (AP) and accuracy. Specifically, SimSiam achieves a higher AP of 75.9 compared to 72.5 and an improved accuracy of 71.2 over the 68.4 seen with the supervised method. However, for the Swin-T backbone, the supervised method demonstrates superior performance with a high AP of 87.9 and an accuracy of 82.2.

Table 2. Linear classification results on industrial target dataset for both ResNet and Swin-T backbones. *ImageNet* refers to the model pretrained with the standard ImageNet dataset, *X-ray SimSiam* denotes the weights obtained from the SimSiam approach pretrained with X-ray images, and *X-ray SimMIM* indicates the weights derived from the SimMIM method under similar conditions.

Backbone	Weights	AP	Accuracy
ResNet	ImageNet	72.5	68.4
	X-ray SimSiam	75.9	71.2
Swin-T	ImageNet	**87.9**	**82.2**
	X-ray SimMIM	70.7	68.6

It is important to note, though, that the results from pretraining with SimMIM are not as satisfactory. This approach appears to highlight the global features of the image rather than the more relevant characteristics associated with defects. Since defects are often quite small relative to the overall image resolution, the features pertinent to these defects tends to be overshadowed by the more prominent structural features of the manufactured product. This tendency tends to diminish the effectiveness in specific defect detection unless fine-tuning is applied to the backbone to accurately teach characteristics of the defects. In such cases, the model learns both the structural features of the product and the local characteristics of the defects, contrasting with the lower AP and accuracy values observed in SimMIM pretraining without such targeted fine-tuning. However, as we explained in Section 4.1, fine-tuning is the optimal protocol for manufacturing defect detection.

4.5. Fine-Tuning on Manufacturing Datasets

Following the linear classification phase, our study progressed into a more comprehensive fine-tuning phase, and we applied the Faster R-CNN model for defect detection on two distinct datasets: our industrial target dataset and the benchmark GDXray dataset. This phase was designed to provide a deeper insight into the practical applicability and effectiveness of the pretrained models in real-world and varied industrial contexts.

In the initial stage of this fine-tuning phase, we trained the Faster R-CNN model on the industrial target dataset using three sets of starting weights: those obtained from X-ray pretraining with both the SimSiam and SimMIM methods and those initialized with ImageNet. This strategy was employed to evaluate the effectiveness of the models in handling complex, real-world defect detection tasks within the manufacturing industry. The results, as detailed in Table 3, illustrate the superior performance of the SimSiam and SimMIM pretraining methods over the conventional supervised approach, especially in identifying critical defects.

Subsequently, our evaluation was extended to the GDXray dataset, replicating the fine-tuning process undertaken with the industrial target dataset but this time applying it to the GDXray training dataset. This step was intended to test the generalizability and robustness of the model in a distinctively different context. The GDXray dataset, with its unique characteristics and challenges, provided an ideal platform for assessing how well the models, pretrained on our industrial unlabeled dataset, could adapt and perform in a broader range of industrial scenarios.

The combined results from both datasets are presented in Table 3, which includes performance metrics on both the industrial target dataset and the GDXray dataset. These results underscore the versatility and efficacy of the SimSiam and SimMIM pretraining approaches in enhancing the performance of defect detection models across diverse X-ray image datasets.

Table 3. Fine-Tuning results on defect detection for manufacturing datasets. The fine-tuning performance of the Faster R-CNN model was evaluated on two datasets, the industrial target dataset and the GDXray dataset, using ResNet and Swin-T backbones with different initial weights (ImageNet, X-ray SimSiam, and X-ray SimMIM). Mean Average Precision (mAP) is reported for the industrial dataset and Average Precision (AP) for the GDXray dataset, designated as (m)AP. This notation is used because the industrial dataset contains two categories of defects, while the GDXray dataset does not categorize defects. Additionally, AP-Critical is reported for the industrial dataset, indicating the accuracy of the model in detecting critical defects.

Dataset	Backbone	Pretraining	(m)AP	AP-Critical
Industrial	ResNet	ImageNet	88.6	94.4
		X-ray SimSiam	89.6	94.8
	Swin-T	ImageNet	**91.3**	94.5
		X-ray SimMIM	**91.3**	**95.5**
GDXray	ResNet	ImageNet	95.7	-
		X-ray SimSiam	**96.0**	-
	Swin-T	ImageNet	94.0	-
		X-ray SimMIM	94.6	-

For the industrial dataset, the performance of the model was quantified using Mean Average Precision (mAP) and AP-Critical, the latter indicating accuracy in detecting critical defects. The results reveal that both SimSiam and SimMIM pretraining methods enhance the performance compared to the traditional ImageNet initialization. Notably, for the ResNet backbone, the SimSiam approach yields a slightly higher mAP of 89.6 and AP-Critical of 94.8 compared to the ImageNet weights. Similarly, the Swin-T backbone pretrained with SimMIM matches the highest mAP of 91.3 achieved by the ImageNet version and even surpasses it in AP-Critical with a score of 95.5.

On the GDXray dataset, the fine-tuning process demonstrated the generalizability of the pretrained models. The SimSiam pretrained ResNet model achieved an AP of 96.0, slightly outperforming the ImageNet-based version from Ferguson et al. [4]. For the Swin-T backbone, the SimMIM pretraining method resulted in an AP of 94.6, showing improved performance over the ImageNet weights. These findings from both datasets underscore the efficacy of the SimSiam and SimMIM pretraining techniques, not only in improving model performance in the specific context of industrial defect detection but also in generalizing well to a broader spectrum of X-ray imaging scenarios.

4.6. Qualitative Results

In this subsection, we present the experimental results through visualizations that highlight the performance of our defect detection model. However, due to confidentiality concerns, we cannot showcase images from the industrial dataset in our visualizations. We instead focus on the GDXray dataset to demonstrate the performance of our defect detection model. Specifically, we highlight the effectiveness of the ResNet backbone pretrained with X-ray images using the SimSiam approach. Through the visualization in Figure 4, we aim to provide insights into the practical application and effectiveness of our model in identifying defects in industrial scenarios.

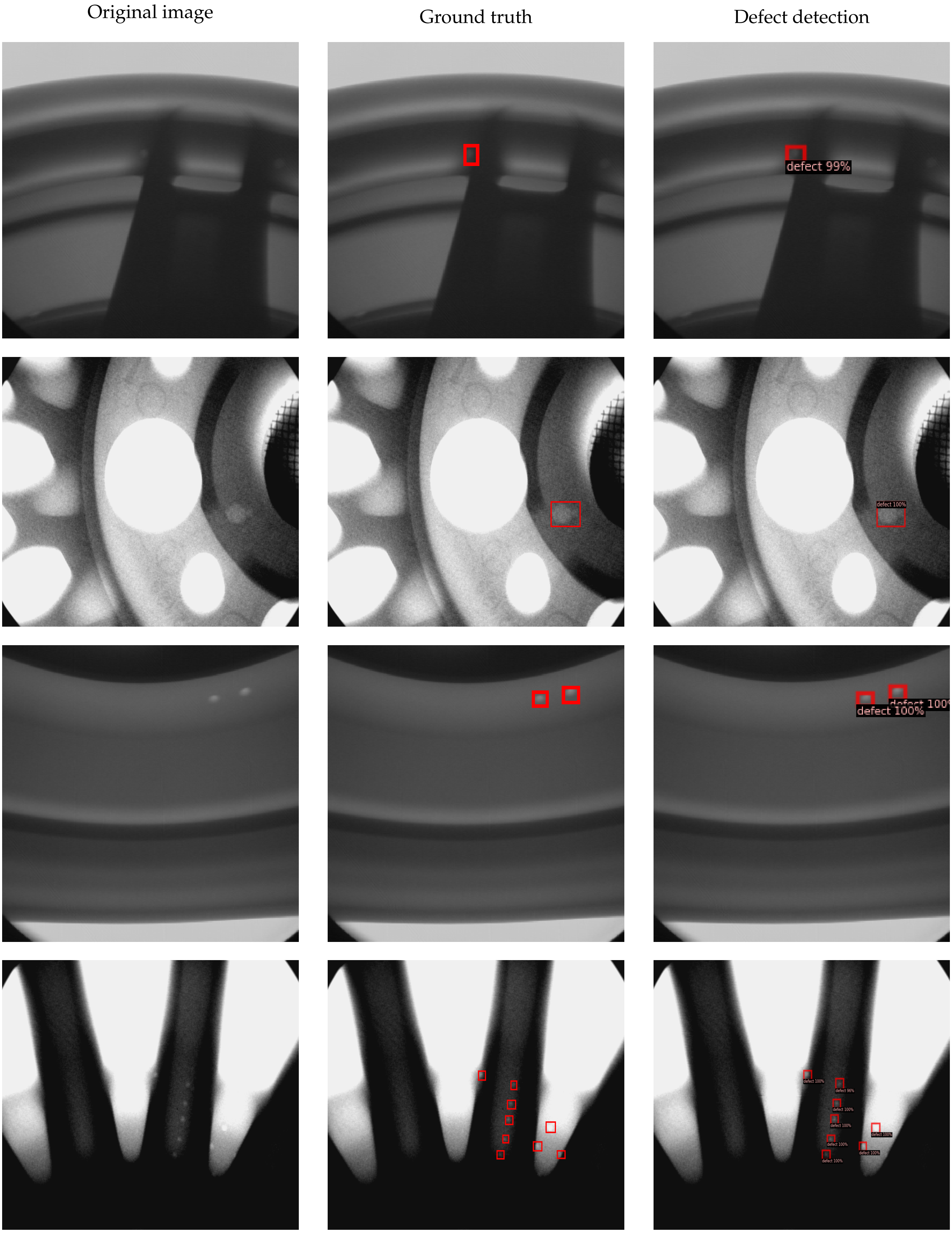

Figure 4. Results of defect detection on the GDXray dataset. This figure illustrates the original X-ray image, ground-truth annotations, and the predictions of the model.

5. Discussion

These experiments shed light on the effectiveness of pretraining methods using X-ray images, especially in the context of detecting defects in manufacturing products. Our experiments show that models pretrained with X-ray images generally surpass those pretrained with ImageNet in discerning relevant features for defect detection. This superiority of domain-specific pretraining highlights the importance of aligning the pretraining phase with the unique characteristics of the task-specific images.

In comparing different backbones, the Swin Transformer models demonstrates superior performance over CNNs when large, labeled data are available [37]. However, the scenario was different with the GDXray dataset, which has fewer labeled images; here, the CNN backbones yielded better results. This suggests that, while Transformers outperform in data-rich environments, CNNs might be more effective with limited data.

A significant aspect of our research is the marked improvement in detecting critical defects following pretraining with X-ray images. This improvement is especially pertinent in industrial settings, where accurately identifying such defects is crucial for safety and reliability. This enhancement in defect detection, particularly for critical defects, could have substantial implications for industrial applications. Moreover, the pretraining on X-ray images demonstrated improved results even with the smaller GDXray dataset, indicating the adaptability of the models to different data volumes. This adaptability is a crucial trait for practical deployment, where data availability can be variable.

While our system has not been implemented in a real production environment, our findings suggest that it holds promise for future applications. The execution time of approximately 0.126 s per image positions our method well within the operational limits set by industry standards, which allows up to a maximum of 1 s for image analysis. This performance metric shows the viability of our method for production line integration. It highlights the potential for real-world deployment, thereby enhancing the feasibility of using deep learning to detect defects in manufacturing settings.

6. Conclusions

This research contributes to the ongoing efforts in enhancing defect detection in manufacturing products using X-ray images. Our comprehensive experiments and analyses show several key findings that underscore the effectiveness and practical applicability of our proposed methodology in industrial settings. Our study conclusively shows that models pretrained on X-ray images consistently outperform those pretrained with ImageNet weights. This finding is crucial as it highlights the importance of domain-specific pretraining in enhancing the ability of the model to discern relevant features for defect detection. By aligning the pretraining phase with the unique characteristics of the task-specific images, we achieved superior detection capabilities compared to those achieved with the models that were pretrained on more general images. Moreover, the comparison of different backbone architectures revealed valuable trends. We observed that, in scenarios with abundant labeled data, Swin Transformer models outperform traditional CNNs, whereas CNN backbones are more effective in datasets with fewer labeled images. This implies that an optimal model can be selected depending on the number of available data and the specific requirements of the task. Particularly noteworthy is the improvement in the detection of critical defects, a crucial concern in industrial settings, achieved through the pretraining on X-ray images. In industrial contexts, where the precise identification of such defects is vital for ensuring safety and maintaining high quality standards, the advancements shown can contribute significantly to reducing risks and enhancing reliability.

In conclusion, this study introduces a methodology that leverages domain-specific pretraining with X-ray images to enhance defect detection in manufacturing products. Our primary contribution lies in the development and application of this methodology, which highlights the advantages of customized pretraining and the strategic choice of a backbone architecture tailored to the specifics of the data. Our methodology significantly improved the accuracy in detecting critical defects, leading to the development of a more robust detection framework. This progress not only advances the frontiers of defect detection technology but also provides valuable insights for the implementation of these models in industrial environments. Such an implementation is likely to enhance the effectiveness of quality control measures, thereby increasing product safety and reliability. Furthermore, the adoption of these enhanced detection systems could lead to a considerable reduction in operational downtime and maintenance expenses, optimize production workflows, and increase overall industrial efficiency. By elevating product quality and minimizing failure rates, our strategy aids in upholding the reputation of manufacturing entities and enhancing consumer confidence.

Author Contributions: Conceptualization, E.I.; Methodology, E.I., C.C. and E.Z.; Software, E.I.; Validation, E.I.; Writing—original draft, E.I.; Writing—review & editing, D.S., C.C. and E.Z.; Visualization, E.I.; Supervision, D.S., C.C. and E.Z.; Project administration, E.Z.; Funding acquisition, C.C. and E.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Eneko intxausti, Carlos Cernuda, and Ekhi Zugasti are part of the Intelligent Systems for Industrial Systems research group of Mondragon Unibertsitatea (IT1676-22), supported by the Department of Education, Universities and Research of the Basque Country. They are also supported by the DREMIND project of the Basque Government under grant KK-2022/00049 from the ELKARTEK program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because of issues related to confidentiality. Furthermore, due to confidentiality agreements, it is not permissible to name the organization that owns the data.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Kuo, J.K.; Wu, J.J.; Huang, P.H.; Cheng, C.Y. Inspection of Sandblasting Defect in Investment Castings by Deep Convolutional Neural Network. *Int. J. Adv. Manuf. Technol.* **2022**, *120*, 2457–2468. [CrossRef]
2. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
3. Ferguson, M.; Ak, R.; Lee, Y.T.T.; Law, K.H. Automatic Localization of Casting Defects with Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 1726–1735. [CrossRef]
4. Ferguson, M.; Ak, R.; Lee, Y.T.T.; Law, K.H. Detection and Segmentation of Manufacturing Defects with Convolutional Neural Networks and Transfer Learning. *Smart Sustain. Manuf. Syst.* **2018**, *2*, 20180033. [CrossRef]
5. Du, W.; Shen, H.; Fu, J.; Zhang, G.; Shi, X.; He, Q. Automated Detection of Defects with Low Semantic Information in X-ray Images Based on Deep Learning. *J. Intell. Manuf.* **2021**, *32*, 141–156. [CrossRef]
6. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
7. Mery, D.; Filbert, D. Automated Flaw Detection in Aluminum Castings Based on the Tracking of Potential Defects in a Radioscopic Image Sequence. *IEEE Trans. Robot. Autom.* **2002**, *18*, 890–901. [CrossRef]
8. Mery, D.; Arteta, C. Automatic Defect Recognition in X-ray Testing Using Computer Vision. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 1026–1035. [CrossRef]
9. Li, X.; Tso, S.K.; Guan, X.P.; Huang, Q. Improving Automatic Detection of Defects in Castings by Applying Wavelet Technique. *IEEE Trans. Ind. Electron.* **2006**, *53*, 1927–1934. [CrossRef]
10. Mery, D. Automated Radioscopic Inspection of Aluminum Die Castings. *Mater. Eval.* **2006**, *65*, 643–647.
11. Tsai, D.M.; Huang, T.Y. Automated Surface Inspection for Statistical Textures. *Image Vis. Comput.* **2003**, *21*, 307–323. [CrossRef]
12. Zhao, X.; He, Z.; Zhang, S.; Liang, D. A Sparse-Representation-Based Robust Inspection System for Hidden Defects Classification in Casting Components. *Neurocomputing* **2015**, *153*, 1–10. [CrossRef]
13. Du, W.; Shen, H.; Fu, J.; Zhang, G.; He, Q. Approaches for Improvement of the X-ray Image Defect Detection of Automobile Casting Aluminum Parts Based on Deep Learning. *NDT Int.* **2019**, *107*, 102144. [CrossRef]
14. Mery, D. Aluminum Casting Inspection Using Deep Object Detection Methods and Simulated Ellipsoidal Defects. *Mach. Vis. Appl.* **2021**, *32*, 72. [CrossRef]

15. Mery, D.; Kaminetzky, A.; Golborne, L.; Figueroa, S.; Saavedra, D. Target Detection by Target Simulation in X-ray Testing. *J. Nondestruct. Eval.* **2022**, *41*, 21. [CrossRef]
16. Parlak, İ.E.; Emel, E. Deep Learning-Based Detection of Aluminum Casting Defects and Their Types. *Eng. Appl. Artif. Intell.* **2023**, *118*, 105636. [CrossRef]
17. Wang, Y.; Hu, C.; Chen, K.; Yin, Z. Self-Attention Guided Model for Defect Detection of Aluminium Alloy Casting on X-ray Image. *Comput. Electr. Eng.* **2020**, *88*, 106821. [CrossRef]
18. García Pérez, A.; Gómez Silva, M.J.; De La Escalera Hueso, A. Automated Defect Recognition of Castings Defects Using Neural Networks. *J. Nondestruct. Eval.* **2022**, *41*, 11. [CrossRef]
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
20. Mery, D.; Riffo, V.; Zscherpel, U.; Mondragón, G.; Lillo, I.; Zuccar, I.; Lobel, H.; Carrasco, M. GDXray: The Database of X-ray Images for Nondestructive Testing. *J. Nondestruct. Eval.* **2015**, *34*, 42. [CrossRef]
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Swizerland, 2014; Volume 8693, pp. 740–755. [CrossRef]
22. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
23. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-Cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
24. Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A.S.; Al-dabbagh, B.S.N.; Fadhel, M.A.; Manoufali, M.; Zhang, J.; Al-Timemy, A.H.; et al. A Survey on Deep Learning Tools Dealing with Data Scarcity: Definitions, Challenges, Solutions, Tips, and Applications. *J. Big Data* **2023**, *10*, 46. [CrossRef]
25. Jing, L.; Tian, Y. Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4037–4058. [CrossRef] [PubMed]
26. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-Supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876. [CrossRef]
27. Van Den Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1747–1756.
28. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 9726–9735. [CrossRef]
29. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1597–1607.
30. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
31. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3733–3742. [CrossRef]
32. Robinson, J.; Chuang, C.Y.; Sra, S.; Jegelka, S. Contrastive Learning with Hard Negative Samples. *arXiv* **2021**, arXiv:2010.04592.
33. Kalantidis, Y.; Sariyildiz, M.B.; Pion, N.; Weinzaepfel, P.; Larlus, D. Hard Negative Mixing for Contrastive Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21798–21809.
34. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap Your Own Latent-a New Approach to Self-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
35. Chen, X.; He, K. Exploring Simple Siamese Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15745–15753. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
37. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
38. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
39. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 6804–6815. [CrossRef]

40. Li, Y.; Wu, C.Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; Feichtenhofer, C. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4794–4804. [CrossRef]
41. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805v2.
42. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
43. Zhou, L.; Liu, H.; Bae, J.; He, J.; Samaras, D.; Prasanna, P. Self Pre-Training with Masked Autoencoders for Medical Image Classification and Segmentation. In Proceedings of the IEEE 20th International Symposium on Biomedical Imaging (ISBI), Cartagena, Colombia, 17–21 April 2023; pp. 1–6.
44. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. SimMIM: A Simple Framework for Masked Image Modeling. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9643–9653. [CrossRef]
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
46. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What Makes for Good Views for Contrastive Learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6827–6839.
47. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin Transformer v2: Scaling up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12009–12019.
48. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
49. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Swizerland, 2015; Volume 9351, pp. 234–241. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.